

S20-82
121957
P-20

N93-15047

Storage Needs in Future Supercomputer Environments

Notes for the presentation by:

Sam Coleman
Lawrence Livermore National Laboratory

July 25, 1991 at the
NASA Goddard "Mass Storage Workshop"

Introduction

The Lawrence Livermore National Laboratory (LLNL) is a Department of Energy contractor, managed by the University of California since 1952. Major projects at the Laboratory include the Strategic Defense Initiative, nuclear weapon design, magnetic and laser fusion, laser isotope separation and weather modeling. The Laboratory employs about 8,000 people. There are two major computer centers: The Livermore Computer Center and the National Energy Research Supercomputer Center.

As we increase the computing capacity of LLNL systems and develop new applications, the need for archival capacity will increase. Rather than quantify that increase, I will discuss the hardware and software architectures that we will need to support advanced applications.

Storage Architectures

The architecture of traditional supercomputer centers, like those at Livermore, include host machines and storage systems linked by a network. Storage nodes consist of storage devices connected to computers that manage those devices. These computers, usually large Amdahl or IBM mainframes, are expensive because they include many I/O channels for high aggregate performance. However, these channels and

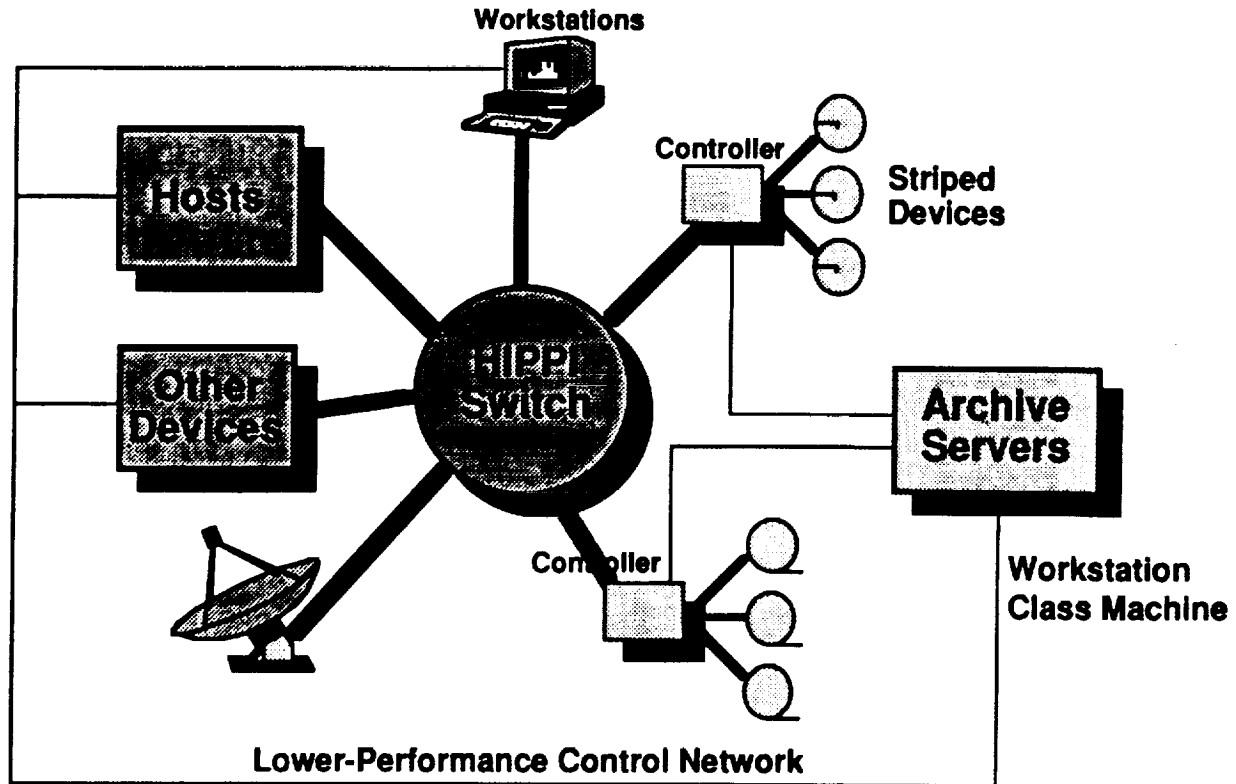
the devices currently attached to them are individually slow; storage systems based on this architecture will become bottlenecks on HIPPI and other high-performance networks. Computers with the I/O-channel performance to match these networks will be even more expensive than the current machines.

The need for higher-performance storage systems is being driven by the remarkable advances in processor and memory technology available on relatively inexpensive workstations; the same technology is making high-performance networks possible. These advances will encourage scientific-visualization projects and other applications capable of generating and absorbing quantities of data that can only be imagined today.

To provide cost-effective, high-performance storage, we need an architecture like that shown in Figure 1. In this example, striped storage devices, connected to a HIPPI network through device controllers, transmit large blocks of data at high speed. Storage system clients send requests over a lower-performance network, like an Ethernet, to a workstation-class machine controlling the storage system. This machine directs the device controllers, also over a lower-performance path, to send data to or from the HIPPI network. Control messages could also be directed over the HIPPI network, but these small messages

would decrease the efficiency of moving large data blocks; since control messages are small, sending them over a slower network will not degrade the overall per-

formance of the system when large data blocks are accessed (this architecture will not be efficient for applications, like NFS, that transmit small data blocks).



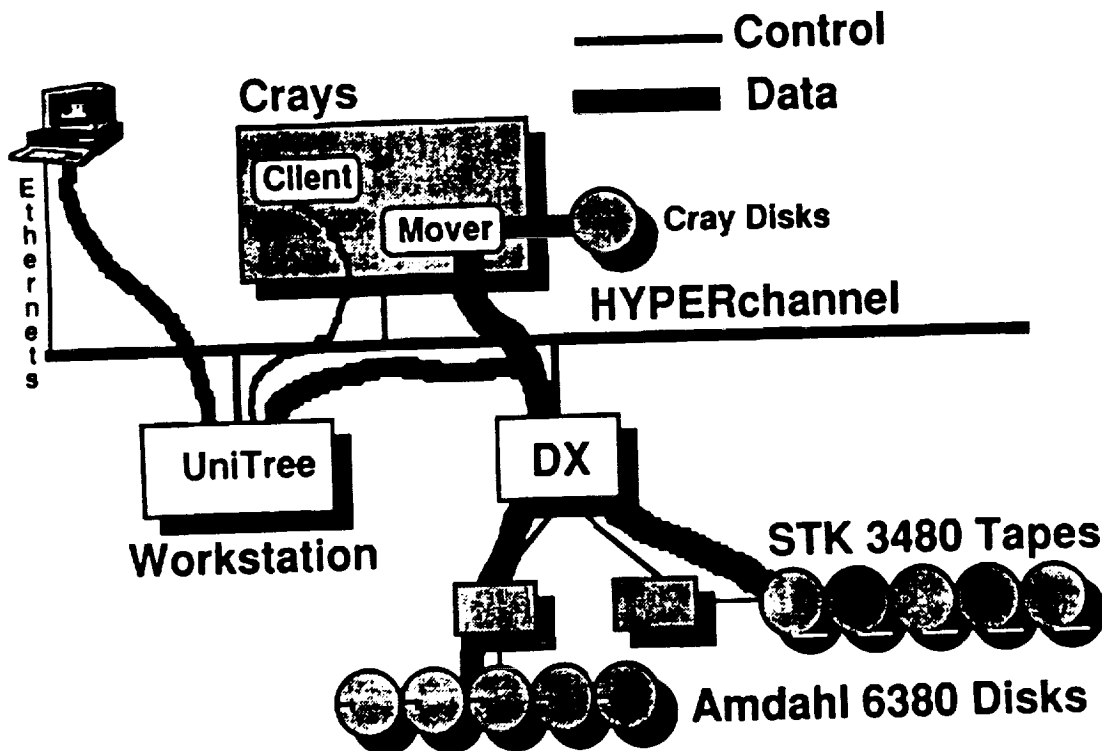
A High-Performance Storage Architecture
Figure 1

To make the architecture in Figure 1 efficient, we will need the following components:

- Programmable device controllers imbedding relatively high-level data-transfer protocols;
- High-performance, possibly striped, archival storage devices to match the performance of the HIPPI network. These devices should be faster than the D1 and D2 magnetic tapes being developed today;
- High-capacity media, with at least the capacity of the largest D2 tape cartridges;
- Robotics to mount volumes quickly;
- Devices and systems that are more reliable than the $1\text{-in-}10^{12}$ error rates quoted today; and
- Devices that are less expensive than the current high-performance devices.

In short, we need reliable, automated archival devices with the capacity of Creo optical tapes (one terabyte per reel), the performance of Maximum Strategy disks (tens of megabytes per second), and the cost of 8mm tape cartridge systems (less than \$100,000).

As a step toward the Figure 1 architecture, we are investigating the architecture shown in Figure 2; we will connect existing storage devices to our Network Systems Corp. HYPERchannel, controlled by a workstation-based UniTree system. Even though the hardware connections are



An Interim Storage Architecture at LLNL
Figure 2

available today, the necessary software is not. In particular, there is no high-level file-transport software in the NSC DX HYPERchannel adapter. As an interim solution, we will put IEEE movers' on our host machines, allowing direct file-transport to and from the storage devices over the HYPERchannel. The UniTree workstation will provide service to client workstations and other network machines. This is acceptable, in the near term, because most of the archival load comes from the larger host machines. This architecture will replace the Amdahl

mainframes that we use to control the current archive.

Software Needs

To implement high-performance storage architectures, we need file-transport software that supports the network-attached devices in Figure 1. Whether or not the TCP/IP and OSI protocols can transmit data at high speeds is subject to debate; if not, we will have to develop new protocols.

From the human client's point of view, we need software systems that provide transparent access to storage. Several transparencies are described in the IEEE Mass Storage System Reference Model document:¹

Access

Clients do not know if objects or services are local or remote.

Concurrency

Clients are not aware that other clients are using services concurrently.

Data representation

Clients are not aware that different data representations are used in different parts of the system.

Execution

Programs can execute in any location without being changed.

Fault

Clients are not aware that certain faults have occurred.

Identity

Services do not make use of the identity of their clients.

Location

Clients do not know where objects or services are located.

Migration

Clients are not aware that services have moved.

Naming

Objects have globally unique names which are independent of resource and accessor location.

Performance

Clients see the same performance regardless of the location of objects and services (this is not always achievable unless the user is willing to slow down local performance).

Replication

Clients do not know if objects or services are replicated, and services do not know if clients are replicated.

Semantic

The behavior of operations is independent of the location of operands and the type of failures that occur.

Syntactic

Clients use the same operations and parameters to access local and remote objects and services.

The IEEE Reference Model

One way to achieve transparency is to develop distributed storage systems that span clients environments. In homogeneous environments, like clusters of Digital Equipment Corp. machines, transparency can be achieved using proprietary software. In more heterogeneous supercomputer centers, standard software, running on a variety of machines, is needed. The IEEE Storage System Standards Working Group is developing standards (project 1244) on which transparent software can be built. These standards will be based on the reference model shown in Figure 3. The modules in the model are:

Application

Normal client applications codes.

Bitfile Client

This module represents the library routines or the system calls that interface the application to the Bitfile Server, the Name Server, and the Mover.

Bitfile Server

The Bitfile Server manages abstract objects called bitfiles that represent uninterpreted strings of bits.

Storage Server

The module that manages the actual storage of bitfiles, allocating media extents, scheduling drives, requesting

volume mounts, and initiating data transfers.

Physical Volume Repository

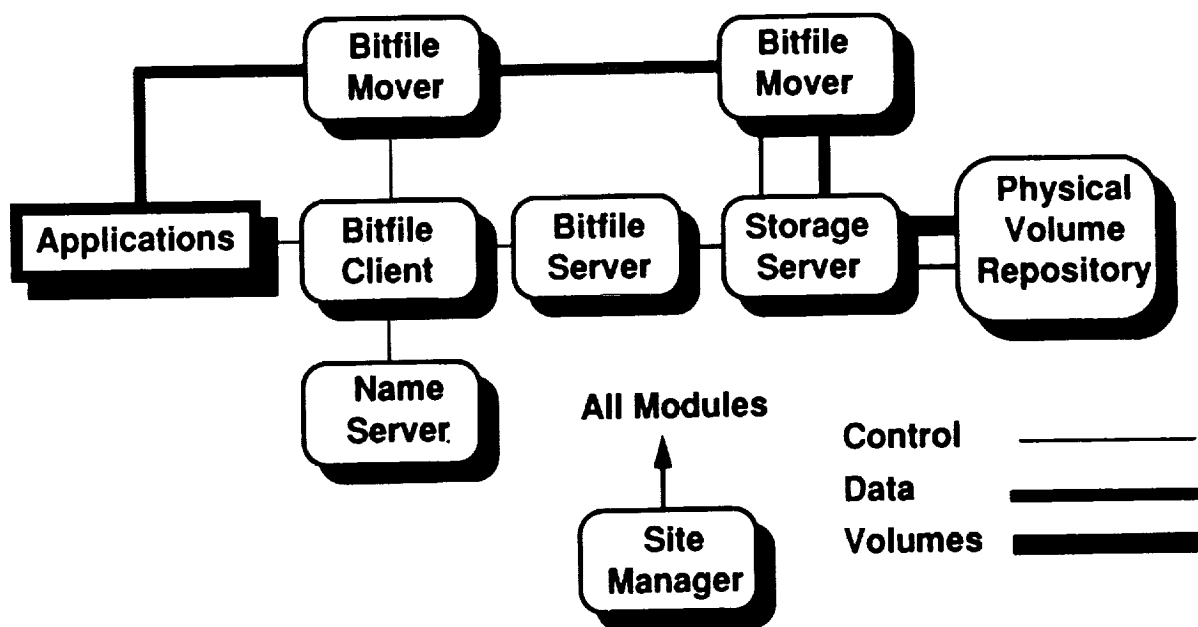
The PVR manages physical volumes (removable disks, magnetic tapes, etc.) and mounts them on drives, robotically or manually, upon request.

Mover

The Mover transmits data between two channels. The channels can be connected to storage devices, host memories, or networks.

Site Manager

This module provides the administration interface to all of the other modules of the model.



The IEEE Mass Storage System Reference Model
Figure 3

The key ideas that will allow standards based on the reference model to support transparency are:

- The Mover separates the data path from the control path, allowing the controller-to-network path shown in Figure 1.
- The Name Server isolates the mapping of human-oriented names to machine-oriented bitfile indenti-

fiers, allowing the other modules in the model to support a variety of different naming environments.

- The modularity of the Bitfile Client, Bitfile Server, Storage Server, and Physical Volume Repository allows support for different devices and client semantics with a minimum of device- or environment-specific software.

I would like to encourage people attending the Goddard conference to support the IEEE standards effort by participating in the Storage System Standards Working Group. For more information, contact me at:

Sam Coleman
Lawrence Livermore National Laboratory
Mail Stop L-60
P. O. Box 808
Livermore, Ca. 94550
(415) 422-4323
scoleman@llnl.gov

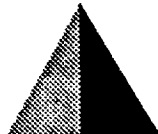
Until standard software systems are available, there are steps that the storage industry can take toward more transparent products. The Sun Microsystems Network File System and the CMU Andrew File System provide a degree of transparency. Work on these systems to improve their security and performance, and to provide links to hierarchical, archival systems, will improve their transparency. I would suggest that software vendors strive to provide operating-system access to archival storage systems, possibly through mechanisms like the AT&T File System Switch.

To learn more about all of the storage issues that I have mentioned, I would encourage you to attend the 11th IEEE Mass Storage Symposium in Monterey, California October 7-10, 1991. For details, contact:

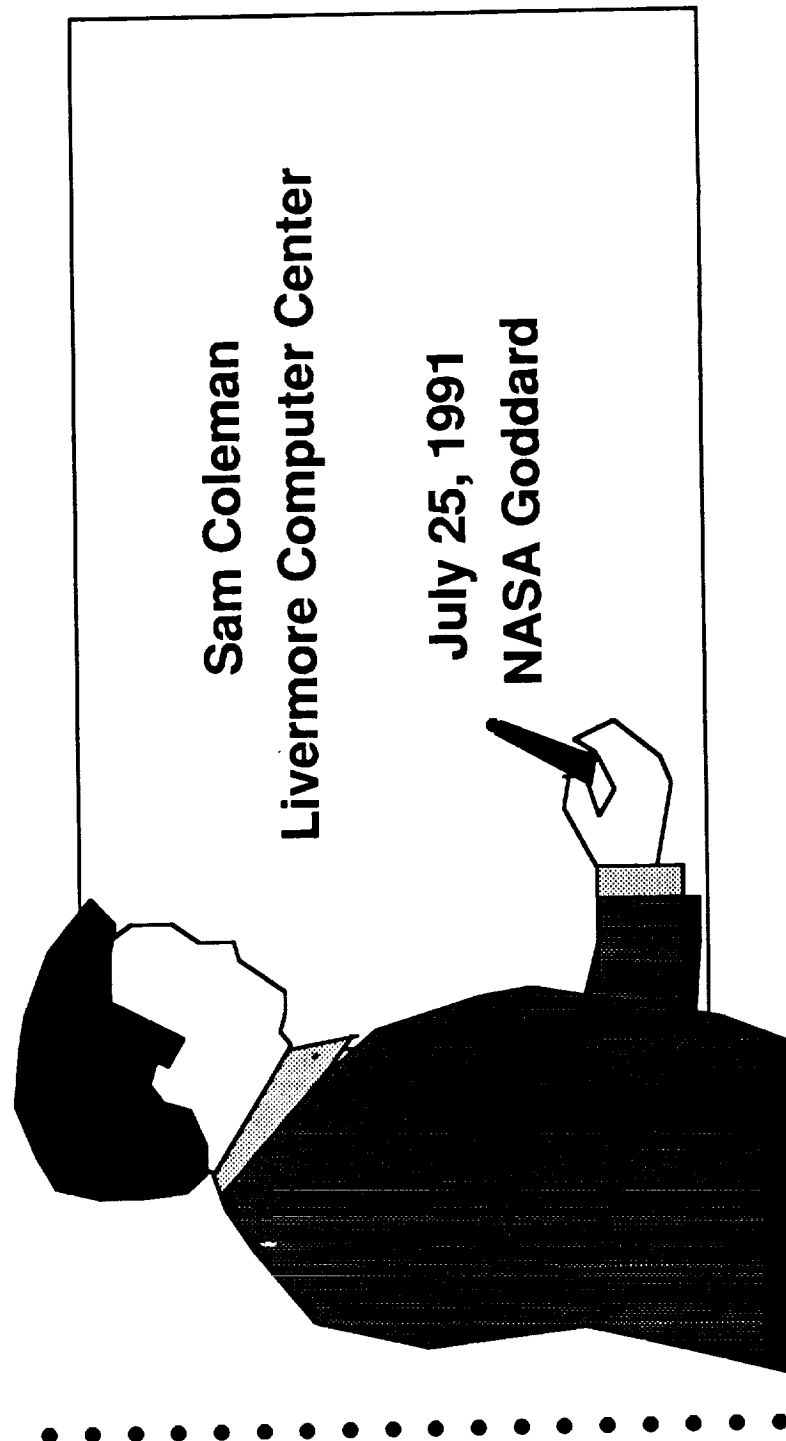
Bernie O'Lear
National Center for Atmospheric
Research
P. O. Box 3000
Boulder, Colorado 80307

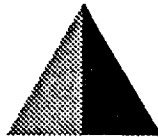
Reference

1. Coleman, S. and Miller, S., editors, *A Reference Model for Mass Storage Systems*, IEEE Technical Committee on Mass Storage Systems and Technology, May, 1990.



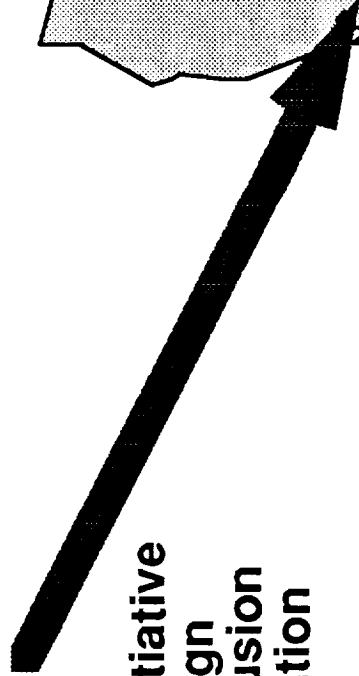
Storage Needs in Future Supercomputer Environments



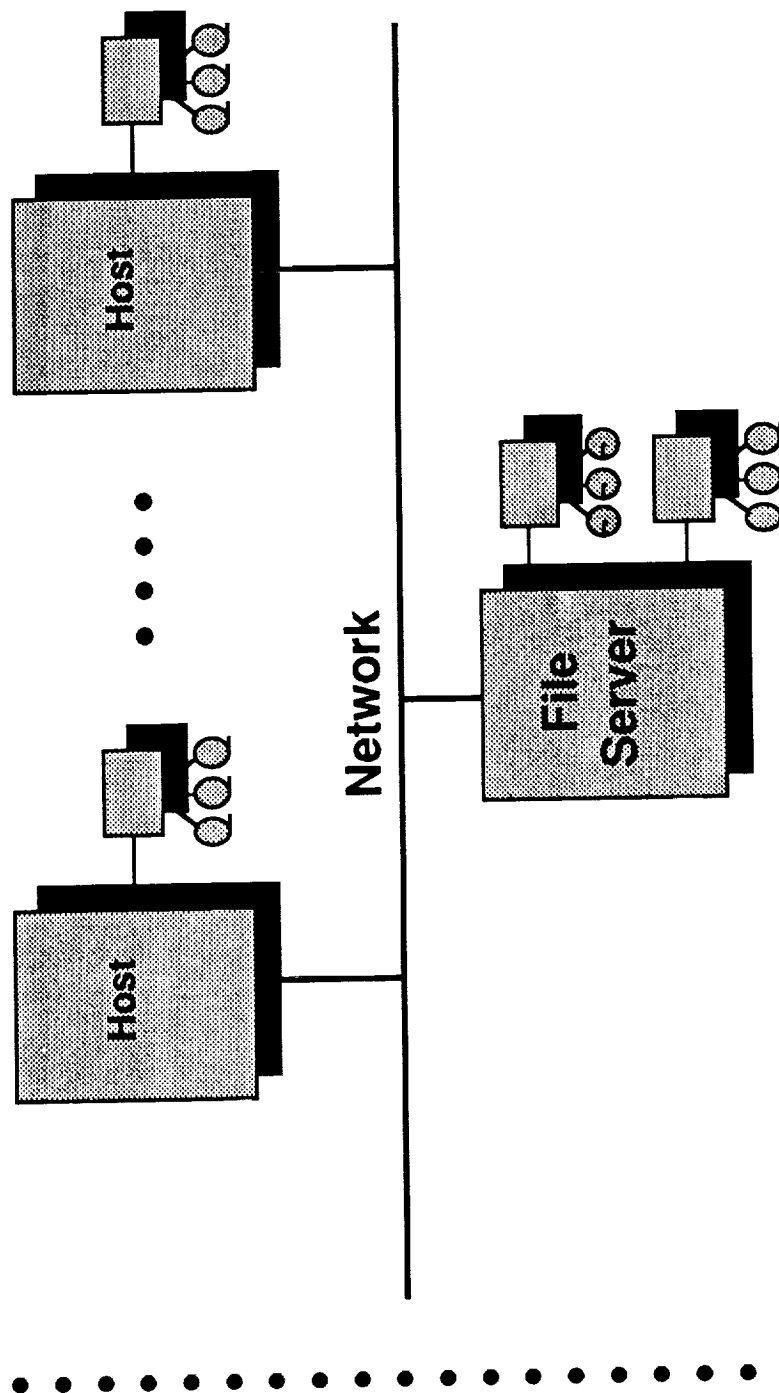


The Lawrence Livermore National Laboratory

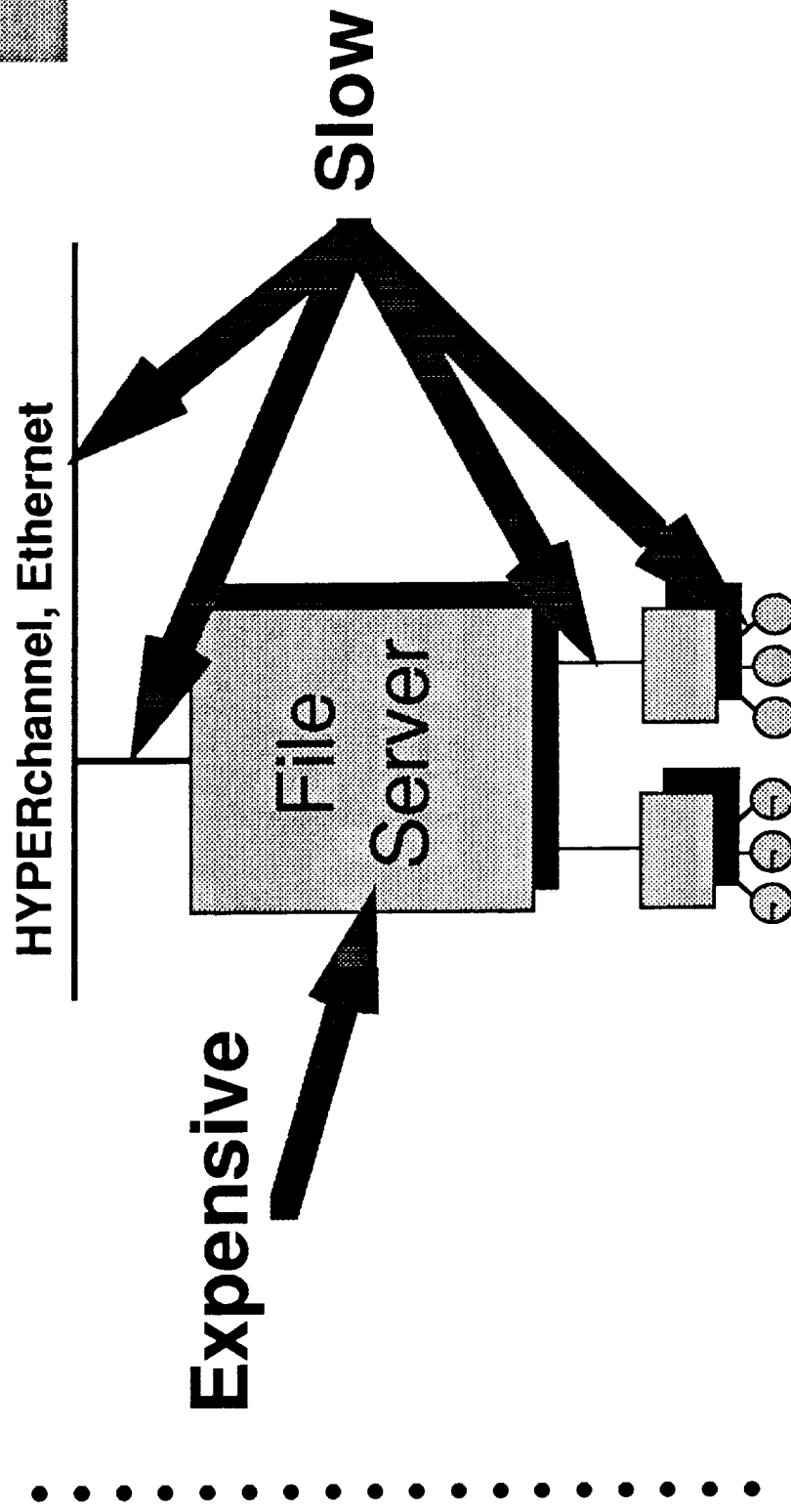
- **Department of Energy contractor**
- **Managed by the University of California**
- **Founded in 1952**
- **Major projects**
 - **Strategic Defense Initiative**
 - **Nuclear weapon design**
 - **Magnetic and laser fusion**
 - **Laser isotope separation**
 - **Weather modeling**
- **8,000 employees, \$1B budget**
- **Two computer centers**
 - **Livermore Computer Center**
 - **National Energy Research Supercomputer Center**



Traditional Supercomputer Storage Architecture



Problems with the Traditional Architecture



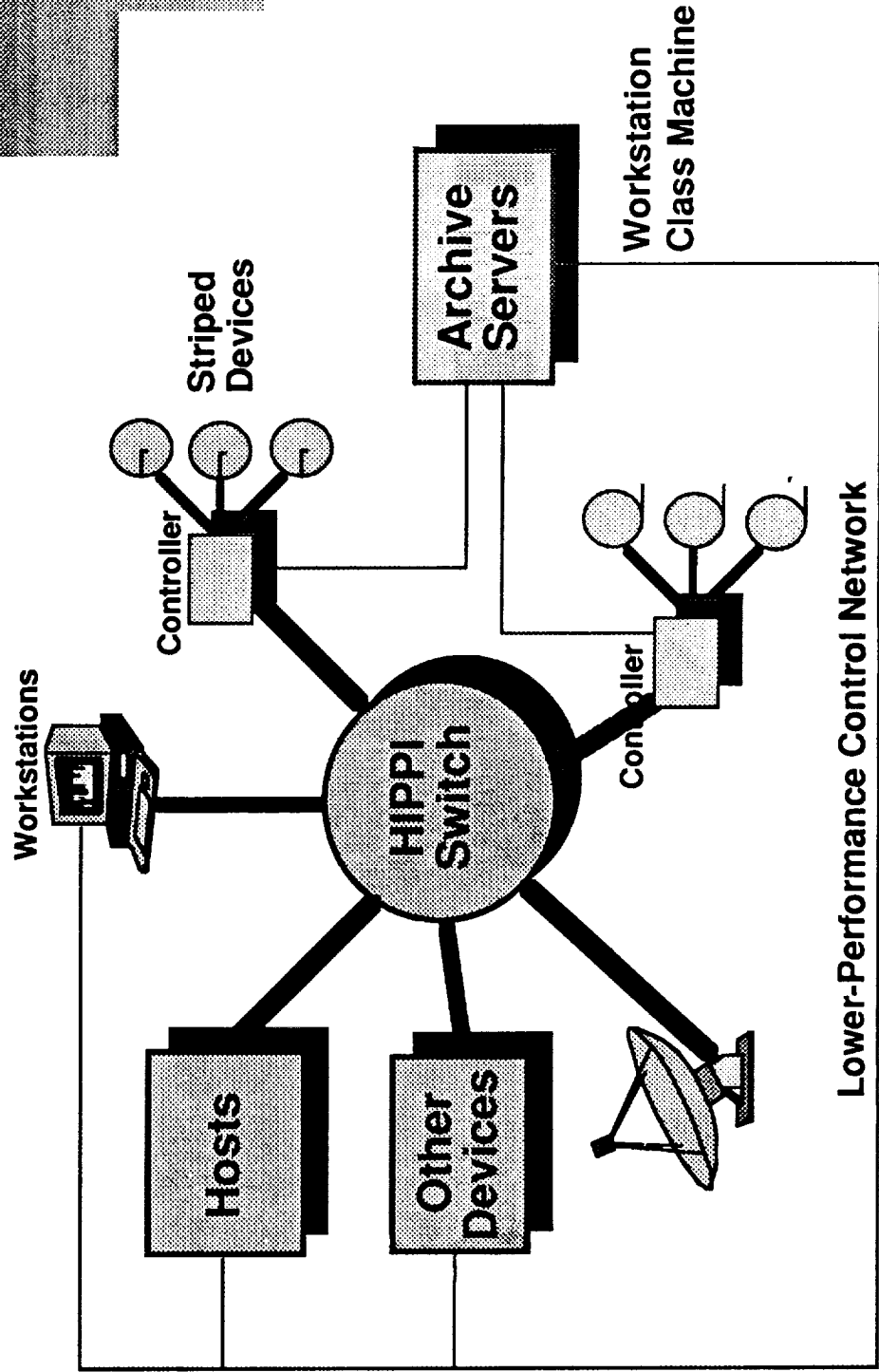


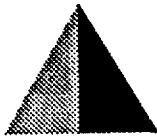
The Need for Higher- Performance Storage

- **Rapidly increasing CPU performance**
- **Exploding main memory sizes**
- **High-performance networks**
- **Scientific visualization**
- **New applications (e.g. Mission to Planet Earth)**
-
-
-

286

A High-Performance Storage Architecture

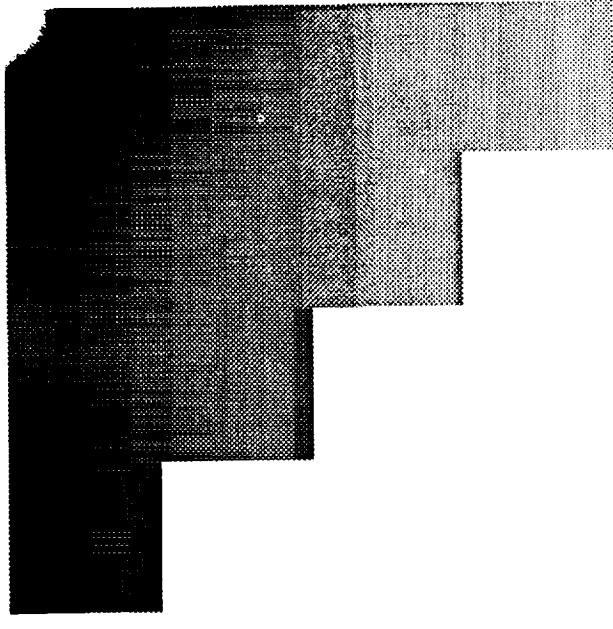


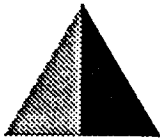


What is Needed

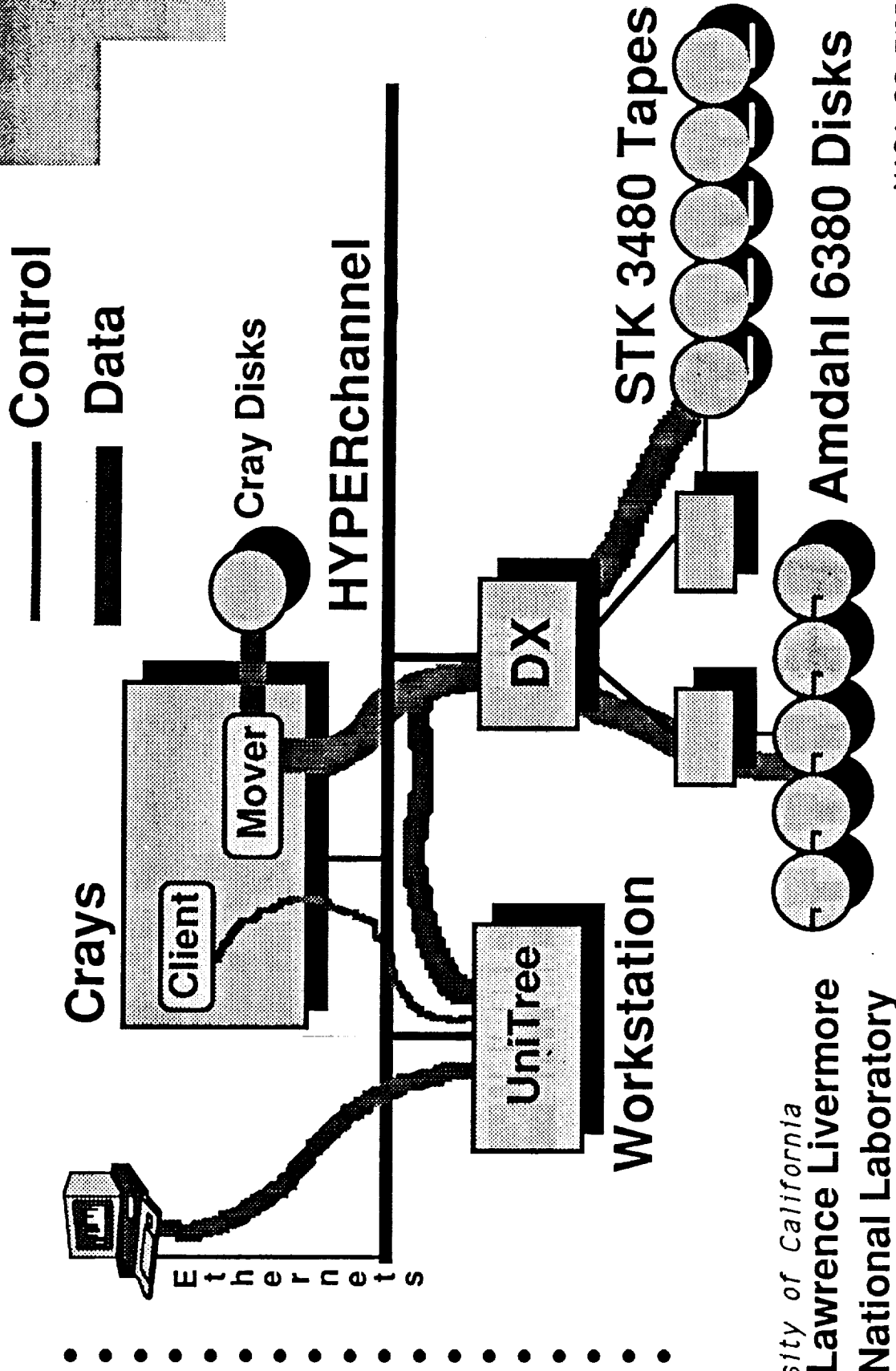
- Programmable device controllers
- For protocols above IPI-3
- Striped devices (RAID)
- HIPPI-speed archival devices
- Faster than D1, D2 tapes
- Striped tapes?
- Higher-capacity media
- Increased reliability
- Cheaper devices, maintenance
-

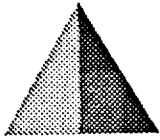
288





In the Meantime.....



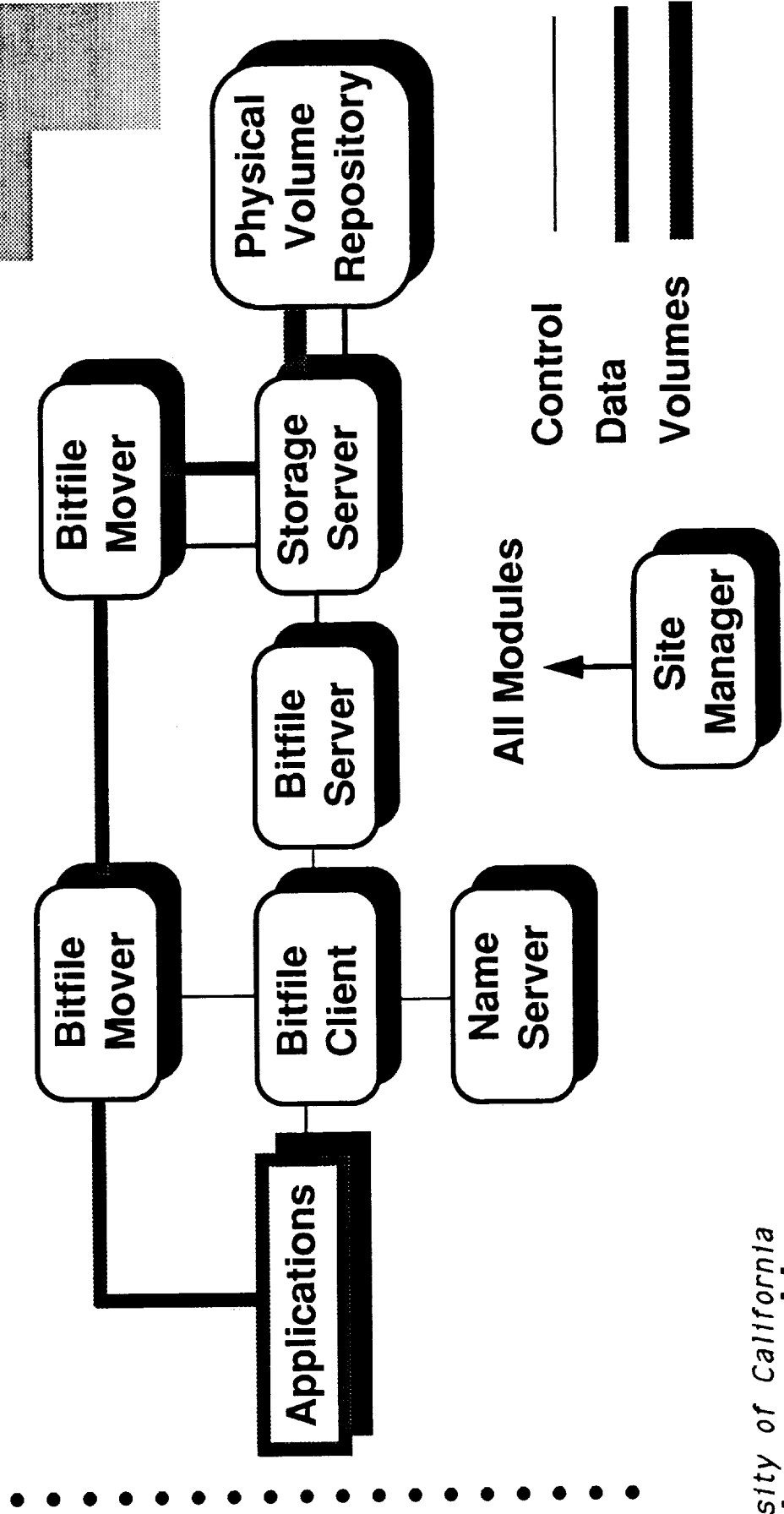


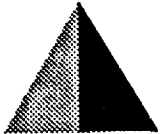
Software Needs

- Support for network-based devices
- Direct data paths
- High performance protocols
- Transparent, distributed systems
- Network-wide naming environments
- Performance transparency
- Device-, location-, operating system-, network-independence
- **Portable, Standard Software!**

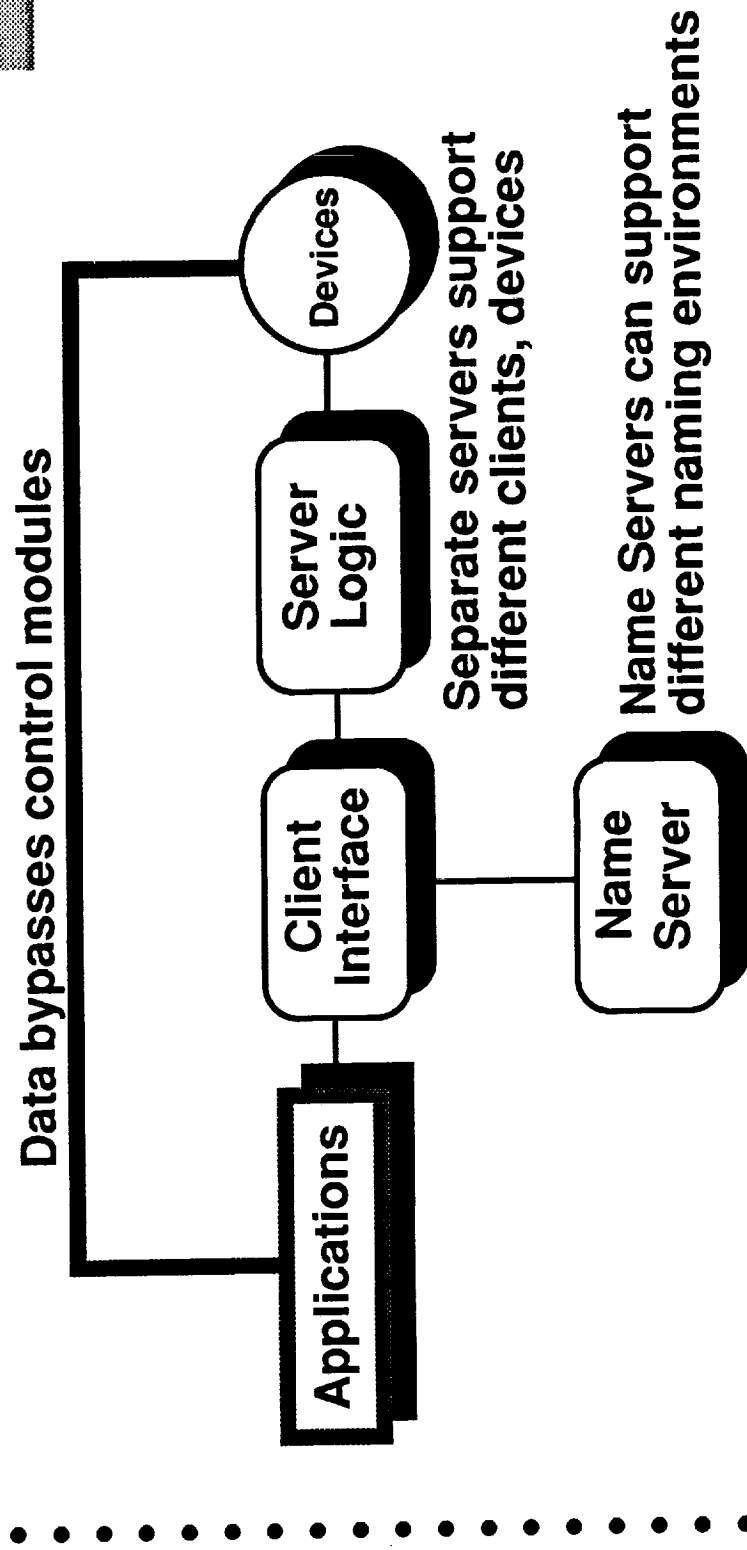
8

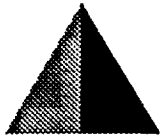
The IEEE Mass Storage System Reference Model





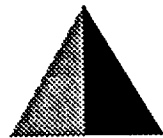
The Significant Modularity





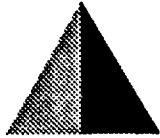
In the Meantime.....

- **We need to go beyond FTP**
- **Sun Network File System**
- **Need to improve security, performance**
- **Andrew File System (AFS, IFS)**
- **Need to integrate with archival systems**
- **File system switch (virtual file system)**
- **Need to provide hierarchical, archival storage**
-
-



Summary of Important Issues for Future Storage Systems

-
-
- **High-performance architectures**
- **Network-attached devices**
- **Device striping technology**
- **Transparent, distributed software architectures**
-
- **Software standards**
- **Open Systems**
-
-
-



To Learn More

- Attend the 11th IEEE Mass Storage
- Symposium
- October 7-10, 1991
- Monterey Sheraton Hotel, Monterey, CA.
- Arranged by
- Bernie O'Lear
- National Center for Atmospheric Research
- P. O. Box 3000
- Boulder, Colorado 80307
-